W2 – ANOVA and DOE

Jeanette Henry

1. Analysis of Variance is used to analyze the general differences between means. It helps an analyst compare the outcomes of the multiple subsets of treatment outcomes used in an experimental study. It is similar to the two-sample t test but can be used on experiments with more than two sets of data. Many articles that I read argued that the Tukey test is more useful and preferred over the ANOVA tests because it gives specific information on which means are different from which. Yet ANOVA is important to learn about in order to understand other people’s data analysis because it is more commonly used.
2. Experimental design is important because the multiple subsets of data in an experiment can be skewed by different variables. Having a well thought out experimental design can limit troublesome factors that could bias your data.

Experimental design relates to ANOVA in that ANOVA is analyzing the means of all your subsets. This could weed out outliers to clean the data, or it could cause the variences in your data to cancel out and possibly go unnoticed.

# One-Way analysis of varience (ANOVA)

https://www.r-bloggers.com/one-way-analysis-of-variance-anova/

In [6]:

plant.df **=** PlantGrowth

In [7]:

summary(plant.df)

weight group   
 Min. :3.590 ctrl:10   
 1st Qu.:4.550 trt1:10   
 Median :5.155 trt2:10   
 Mean :5.073   
 3rd Qu.:5.530   
 Max. :6.310

In [8]:

plant.df**$**group **=** factor(plant.df**$**group,

labels =c("Control", "Treatment 1", "Treatment 2"))

In [9]:

require(ggplot2)

Loading required package: ggplot2  
Warning message:  
"package 'ggplot2' was built under R version 3.3.3"

In [14]:

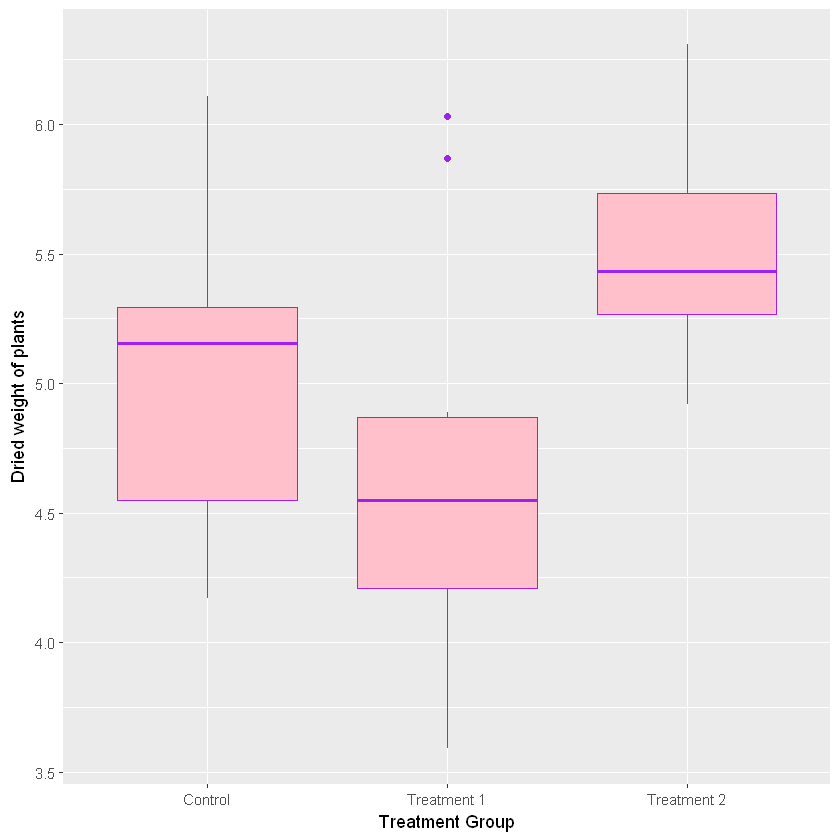
ggplot(plant.df, aes(x = group, y = weight)) **+**

geom\_boxplot(fill = "pink", colour = "purple") **+**

scale\_x\_discrete() **+** xlab("Treatment Group") **+**

ylab("Dried weight of plants")

*#This is a boxplot of the dried weights for the three competing groups*



In [ ]:

*#We visually see that there are different, however using ANOVA we can quantify it.*

In [15]:

plant.mod1 **=** lm(weight **~** group, data = plant.df)

summary(plant.mod1)

Call:  
lm(formula = weight ~ group, data = plant.df)

Residuals:  
 Min 1Q Median 3Q Max   
-1.0710 -0.4180 -0.0060 0.2627 1.3690

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.0320 0.1971 25.527 <2e-16 \*\*\*  
groupTreatment 1 -0.3710 0.2788 -1.331 0.1944   
groupTreatment 2 0.4940 0.2788 1.772 0.0877 .   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom  
Multiple R-squared: 0.2641,        Adjusted R-squared: 0.2096   
F-statistic: 4.846 on 2 and 27 DF, p-value: 0.01591

(Intercept) refers to the first group, which is the control group.

Under the Estimate column, we see the mean of the control group is 5.0320 and down the column is the differences of the control group's mean and the means of the test groups.

THis gives us specifice information about coeeficients.

Last column is a p-value for the difference of that mean from zero. Treatment 2 has a significant difference from the control group.

In [16]:

anova(plant.mod1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
| **group** | 2 | 3.76634 | 1.8831700 | 4.846088 | 0.01590996 |
| **Residuals** | 27 | 10.49209 | 0.3885959 | NA | NA |

Runs f test on the factor of the treatment we have a p value of .0159. Because this is a single factor model the p-value in anova matches the p-value in lm. THis p-value still shows that there are significant differences between the groups.

In [17]:

confint(plant.mod1)

|  |  |  |
| --- | --- | --- |
|  | **2.5 %** | **97.5 %** |
| **(Intercept)** | 4.62752600 | 5.4364740 |
| **groupTreatment 1** | -0.94301261 | 0.2010126 |
| **groupTreatment 2** | -0.07801261 | 1.0660126 |

These are the confidence intervals from the logistic regression model

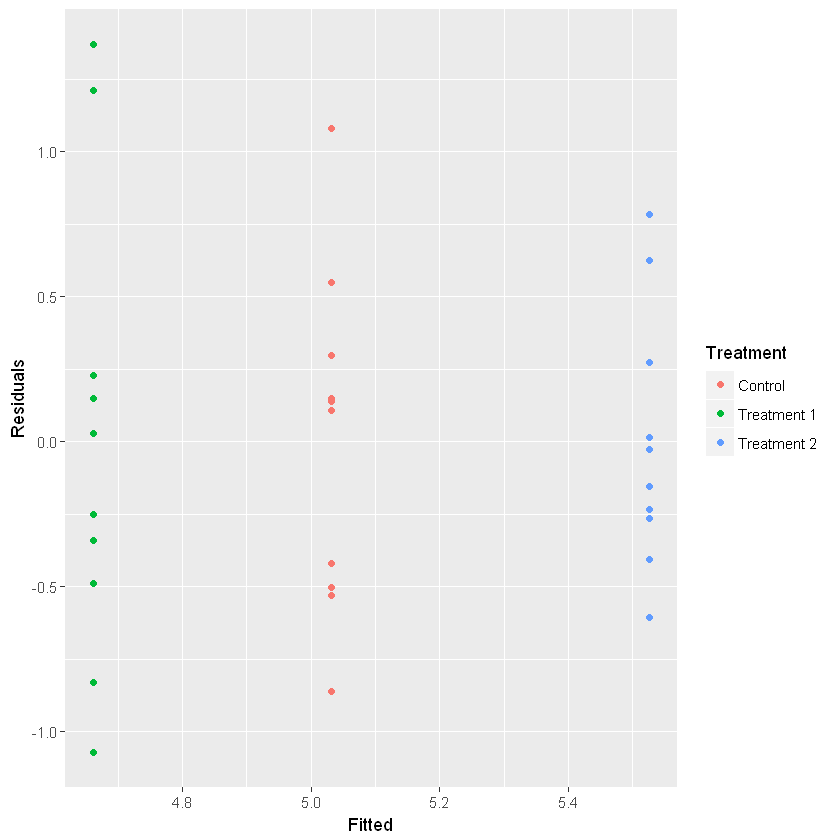
In [18]:

plant.mod **=** data.frame(Fitted = fitted(plant.mod1),

Residuals = resid(plant.mod1), Treatment = plant.df**$**group)

In [19]:

ggplot(plant.mod, aes(Fitted, Residuals, colour = Treatment)) **+** geom\_point()



In [ ]:

Diagnostic plot does not have major problems, but we do see variability in the spread of residuals for the 3 groups.